

Jiří Zváček: Aplikace částečné a segmentové regrese.

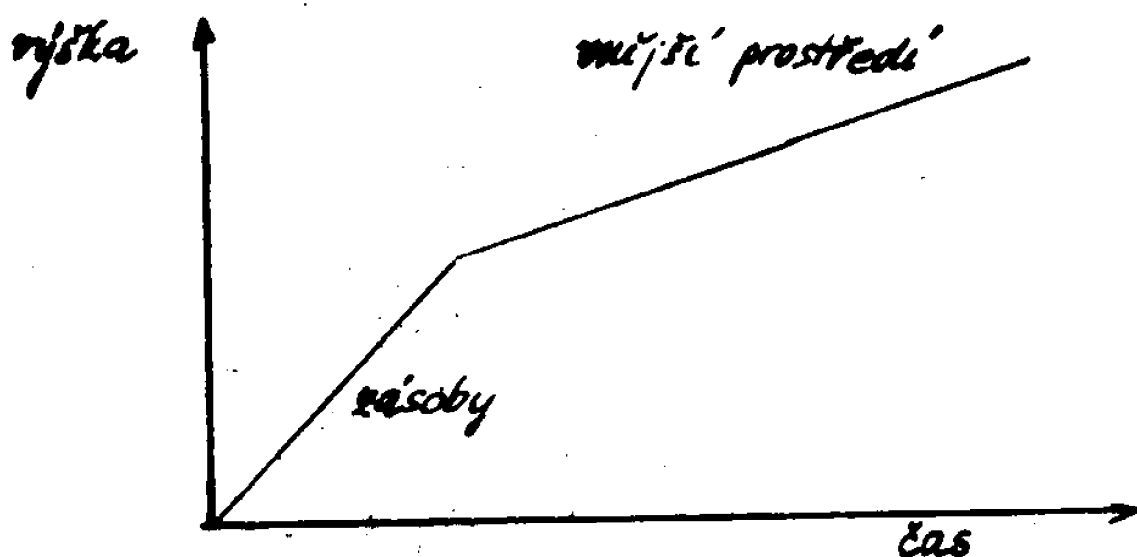
V této práci je dán určitý přehled metod používaných v částečné a segmentové regresi a příklad praktické aplikace. Cílem je dát čtenáři úvodní představu o této málo známé a přitom velmi perspektivní problematice, ve které je ještě dosti prostoru i pro samostatnou práci. Vzhledem k rozsáhlosti tématu na jedné straně a omezenosti, zejména místa, na straně druhé, jsou problémy i popisy řešení spíše naznačeny, než podrobně a přesně popsány. Podrobnosti nalezneme čtenář v citované literatuře.

1. Úvod a terminologie.

V klasické regresní analýze, v regresní analýze časových řad a v ekonometrii se předpokládá, že regresní model, používaný pro popis závislosti mezi pozorováními je platný pro všechna pozorování, která máme k dispozici a že platí i pro případná pozorování další /tedy i pro interpolační a extrapolaci úlohu/.

Tento předpoklad je v řadě situací neudržitelný. Klasický příklad je z oblasti biologie, kde zřejmě musíme předpokládat zcela jiný model pro dynamiku vývoje rostliny v čase v té době, když spotřebovává živiny ze zásob /semene/ a jiný model v době, kdy je ovlivněna na okolí. Přitom na základě pozorování dynamiky vývoje rostliny v čase máme právě bod, ve kterém dochází k tomuto kvalitativnímu slomu, určit.

Graficky je tato situace zobrazena na obrázku číslo 1.



Obrázek č. 1.

Příkladů tohoto typu lze nalézt celou řadu. Je zřejmé že je principiálně nemožné popsat takovouto heterogenní situaci jedním modelem, platným pro všechna pozorování. Přinejmenším je jednodušší modelovat každou z těchto situací zvlášť.

Místo klasického regresního modelu

$$(1) \quad E(y_i | x_i) = f(x_i, \beta) \quad , i=1,2,\dots,n \quad ,$$

kde y_i jsou pozorování závislé proměnné,
 x_i je vektor pozorování nezávisle proměnné
 β je vektor neznámých parametrů,
 n je počet pozorování a
 f je známá funkce,

pracujeme tedy s regresním modelem "rozděleným" na oblasti M_j

$$(2) \quad E(y_i | x_i) = f_j(x_i, \beta_j) \quad \text{pro } x_i \in M_j \\ j=1,2,\dots,r \quad ,$$

kde r je počet oblastí /často neznámý/,
 f_j jsou známé funkce,
 β_j jsou neznámé vektory parametrů,
 M_j je j -tá oblast vysvětlujících proměnných
/obecně též neznámá/.

Je přirozené, že tento obecný model má celou řadu variant podle toho, jaká další omezení klademe na rozdělený model.

V oblasti ekonomie je velmi častým omezením požadavek stálosti regresní funkce. Toto omezení sice není podstatné a většina výsledků je platná i bez tohoto omezení, pro jednoduchost symboliky jej však přijmeme.

Platí tedy

$$(3) \quad f_j \equiv f \quad , \text{pro } j=1,2,\dots,r \quad .$$

Nejdůležitější aplikace pocházejí z oblasti časových řad. V regresní analýze časových řad je vektor vysvětlujících proměnných závislý pouze na časové proměnné t , takže model má tvar

$$(4) \quad r(x_1 | \beta_j) = r(t, \beta_j) \quad \text{pro } t=1,2,\dots,T$$

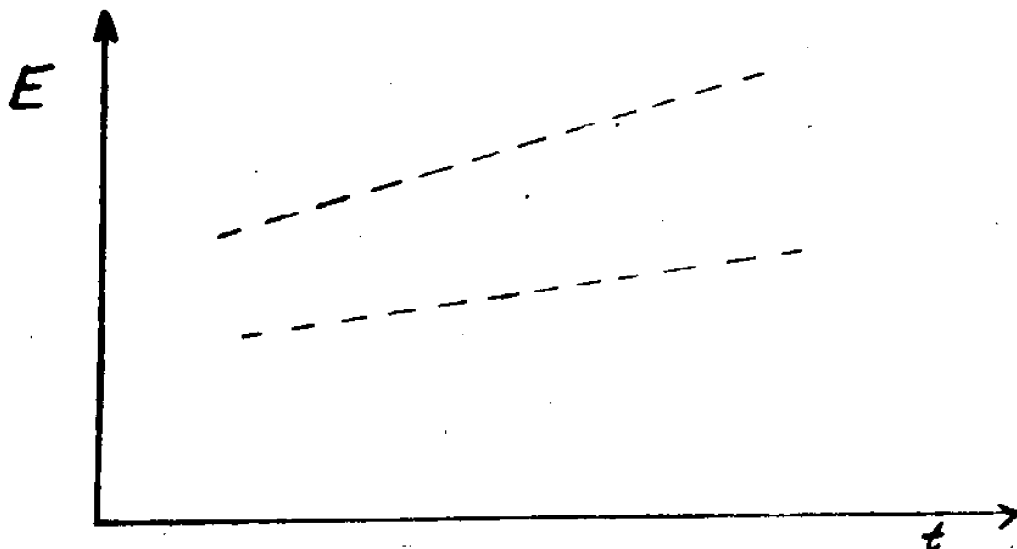
a obecněji v ekonometrických modelech dostaneme pro jednotlivé rovnice závislost na vektoru vysvětlujících proměnných X v časové okamžiku t , takže model můžeme zapsat jako

$$(5) \quad r(x_t | \beta_j) = r(x_t, \beta_j) \quad \text{pro } t=1,2,\dots,T.$$

Oblasti M_j se tedy liší pouze v parametrech regresní funkce. Modely v této podobě jsou velmi podobné tzv. splisové regresi. Je však třeba si uvědomit, že je zde přinejmenším jeden podstatný rozdíl - oblasti M_j zde nesaháme. Další rozdíly pak vyplývají z toho, že regresní model zde má i věcnou interpretaci - nejedná se nám o pouhou aproximaci.

Je zřejmé, že různé varianty této rozdělené úlohy se budou lišit zejména podle toho, jaké předpoklady vyslovíme o charakteru oblastí M_j .

Pokud nevyslovíme žádné další předpoklady o oblastech M_j , hovoříme o přeskokové /switching/ regresi. Prakticky se jedná v tomto případě o směs různých modelů. Situace je naznačena na obrázku čísla 2. Tento případ není prakticky příliš zajímavý, neboť o oblastech M_j se "dozvíme" pouze to, kolik procent pozorování do nich patří.



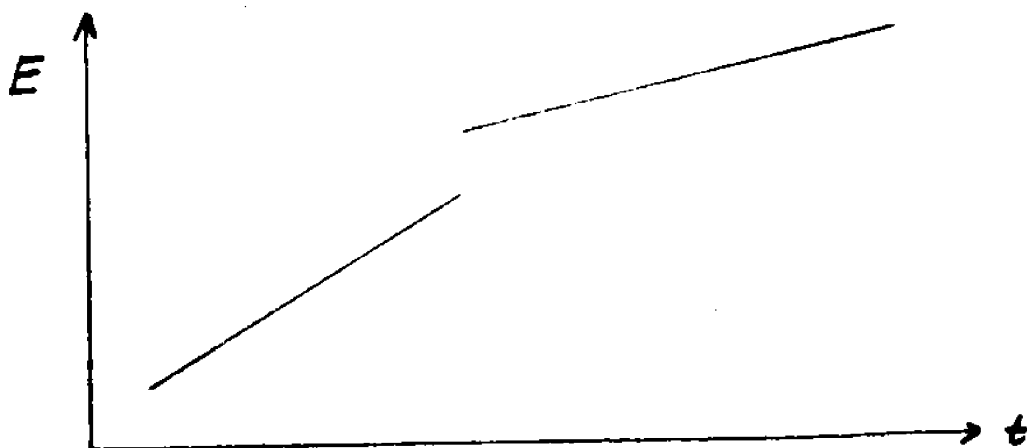
Obrázek č. 2.

Pro praxi jsou zajímavější modely, ve kterých jsou oblasti M_j uzavřené. Tak je tomu zejména v regresní analýze časových řad, kde je oblast M_j dána intervalem časové proměnné

$$(5) \quad M_j = \langle t_{j-1}, t_j \rangle \quad , j=1, 2, \dots, r$$

$$\text{kde } t_0 = 1 \text{ a } t_r = T .$$

V tomto případě hovoříme o částečné /piecewise/ regresi. Tato situace je schematicky znázorněna na obrázku číslo 3. Body t_j , které určují oblasti M_j nazýváme dělicí body.



Obrázek č. 3

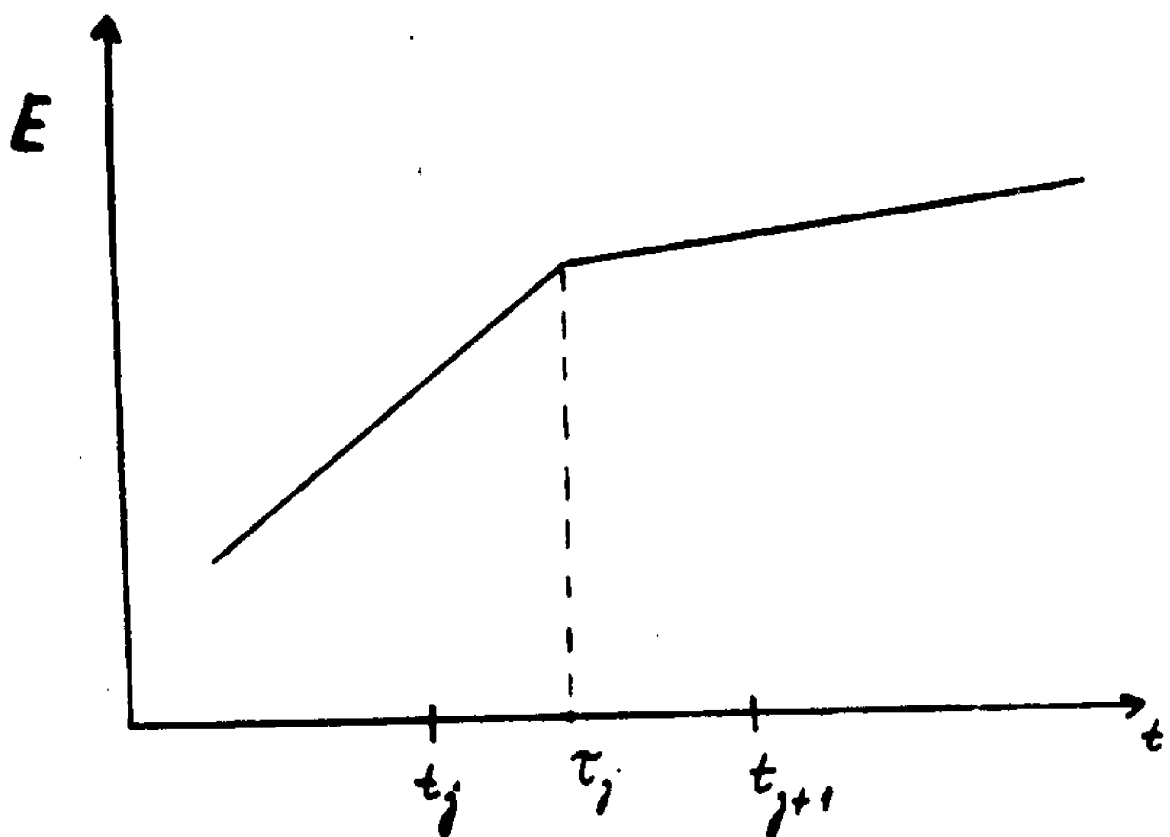
Dalším omezením je omezení na spojitost. Na rozdíl od předchozího případu částečné regrese zde musí navíc platit

$$(6) \quad r(\tau_j, \beta_j) = r(\tau_j, \beta_{j+1}) \quad , j=1, 2, \dots, r$$

$$\text{pro } t_j \leq \tau_j \leq t_{j+1} .$$

V některém časovém okamžiku τ_j tedy jeden model přejde spojitě v druhý /viz obrázek číslo 4/. Dělicí bod τ_j zde obecně nemusí být totožný s některým pozorováním.

Této úloze říkáme segmentová regrese. Je zřejmé, že náš úvodní příklad je právě úlohou segmentové regrese.



Obrázek č. 4

Jednotlivé varianty segmentové regrese se liší podle charakteru dělicích bodů τ_j , kde je nutno rozlišit dva případy

$$(7) \quad \tau_j = t_j \quad , j=1, 2, \dots, r \quad .$$

tedy když k přechodu dochází pouze v pozorovaných bodech, a

$$(8) \quad t_j < \tau_j < t_{j+1} \quad , j=1, 2, \dots, r \quad .$$

Často se vyslovuje i další předpoklad, o hladkosti přechodu regresního modelu

$$(9) \quad f'(\tau_j, \beta_j) = f'(\tau_j, \beta_{j+1})$$

kdy tedy i /první/ derivace musí být v dělicím bodě τ_j spojitá.

Prakticky nejzajímavější úlohy se vyskytují v částečné a segmentové regresi. V této přehledové práci se speciálně zaměříme na částečnou regresi, neboť je typická právě pro oblast ekonomie, kde změny mají charakter skoků /cenových, v dynamice atd. /.

Literatura v této oblasti je překvapivě rozsáhlá, ale bohužel i velmi pestrá a s nejednotnou terminologií. Přehledná monografie, která by postihovala celou problematiku z jediného hlediska chybí.

Z rozsáhlejších prací lze doporučit práci HACKL /1980/, která je však věnována spíše otázkám testování hypotézy o stabilitě modelu a zejména v ruském překladu dostupnou práci POIRIER /1976/, kde je této problematice věnována přehledová 7. kapitola.

Vybrané články nalezneme čtenář v seznamu literatury, podle obsahu se pak na ně odvoláváme v textu.

2. Odhady parametrů lineárního modelu v částečné regresi.

Pro ilustraci cesty, která vede ke konstrukci kritéria odhadu parametrů a dělicích bodů, uvedeme postup Quandtova řešení nejjednoduššího případu klasického lineárního regresního modelu s jediným dělicím bodem τ . Na něm jsou zároveň vidět i slabiny tohoto přístupu, který nelze nazvat po všech stránkách exaktním.

Model, zkoumaný např. v pracích QUANDT /1958, 1960/ je

$$(10) \quad y^{(j)} = X^{(j)} \beta^{(j)} + \varepsilon^{(j)} \quad , j=1,2$$

$$\begin{aligned} \text{oblast } M_1 &= (1, 2, \dots, \tau) \\ M_2 &= (\tau+1, \dots, T) \end{aligned}$$

Náhodné poruchy ε_t jsou vzájemně nezávislé, s nulovou střední hodnotou a konstantním rozptylem v každé oblasti /heteroskedastické/ a mají normální rozdělení.

Platí tedy

$$(11) \quad \varepsilon^{(j)} \sim N[0, \sigma_j^2 I_j] .$$

Logaritmus věrohodnostní funkce tohoto modelu $l(\tau)$ lze tedy zapsat ve formě

$$(12) \quad l(\tau) = -T \ln \sqrt{2\pi} - \tau \ln \sigma_1^2 - (T - \tau) \ln \sigma_2^2 - \\ - \sum_{j=1}^2 \frac{1}{2\sigma_j^2} \left[y^{(j)} - X^{(j)} \beta^{(j)} \right]^2$$

kde jsme použili zkráceného označení $[A] = A'A$.

Quantitativně navrhuje minimalizovat věrohodnostní funkci (12) ve dvou krocích.

V prvním kroku konstruuje tzv. lokální odhady nejmenších čtverců pro odhady vektorů parametrů $\beta^{(j)}$ a rozptylů σ_j^2 , které jsou

$$\hat{\beta}^{(j)} = b^{(j)} = \left[X^{(j)'} \right]^{-2} X^{(j)'} y^{(j)}, \quad j = 1, 2 \\ (13) \quad \hat{\sigma}_1^2 = s_1^2 = \frac{1}{\tau} \left[y^{(1)} - X^{(1)} b^{(1)} \right]^2, \\ \hat{\sigma}_2^2 = s_2^2 = \frac{1}{T - \tau} \left[y^{(2)} - X^{(2)} b^{(2)} \right]^2.$$

Odhad dělicího bodu τ získáme ve druhém /iterativním/ kroku maximalizací $l(\tau)$, kde za σ_j^2 dosadíme odhady s_j^2

$$(14) \quad l(\tau) = -T \ln \sqrt{2\pi} - \frac{\tau}{2} \ln s_1^2 - \frac{T - \tau}{2} \ln s_2^2 - \frac{T}{2}.$$

Maximálně věrohodný odhad parametrů tedy dostaneme minimalizací výrazu

$$(15) \quad Q = \ln s_1^2 + (T - \tau) \ln s_2^2.$$

Analogickým postupem dostaneme odhady pro homoskedastické náhodné poruchy, kde předpokládáme, že rozptyl náhodných poruch je /všude/ konstantní. Ostatní předpoklady zůstávají stejné.

Platí tedy

$$(16) \quad \varepsilon^{(j)} \sim N[0, \sigma^2 \Pi_j], \quad j=1, 2.$$

Při použití lokálních odhadů nejmenších čtverců dostáváme tedy dva /různé/ odhady pro rozptyl σ^2 /viz rovnice (13) /. Jako odhad rozptylu σ^2 použijeme tzv. sprůměrovaný /pooled/ odhad

$$(17) \quad \hat{\sigma}^2 = s^2 = \frac{\tau s_1^2 + (\tau - \tau) s_2^2}{\tau},$$

takže maximálně věrohodný odhad parametru τ dostaneme /analogicky/ minimalizací výrazu

$$(18) \quad Q = \tau s^2 + (\tau - \tau) s^2 = \tau s^2,$$

což není nic jiného než součet čtverců reziduí v obou oblastech.

Obecně tedy pro r oblastí dostaneme odhady tohoto typu /viz např. HAWKINS (1976) / pro heteroskedastické modely minimalizací výrazu

$$(19) \quad Q^{(r)} = \sum_{j=1}^r m_j \ln s_j^2,$$

kde m_j je počet pozorování v oblasti M_j ,

a pro homoskedastické modely minimalizací výrazu

$$(20) \quad Q^{(r)} = \sum_{j=1}^r m_j s_j^2 = \sum_{j=1}^r Q_j,$$

kde Q_j je lokální součet čtverců reziduí.

Uvedená metoda je vhodná pro odhad při zadaném počtu oblastí. Problém vyhledávání r nelze takto řešit a budeme se mu věnovat v části věnované testování hypotéz a v aplikační části.

V praktických situacích dáme většinou přednost homoskedastické formulaci úlohy. Vyplyvá to například z toho, že oblasti M_j musíme v heteroskedastických úlohách omezit na případy, kdy $n_j > p$, kde p je počet parametrů jednotlivých regresních modelů /pro $n_j = p$ nám totiž vyjde odhad $\hat{\sigma}_j^2$ nulový / a i z jiných nepříjemných vlastností logaritmu.

Pro praktické úlohy je uvedené řešení postačující, z teoretického hlediska je však se dvou důvodů problematické.

Pro případ homoskedastických modelů není toto řešení exaktní. Ve skutečnosti se jedná o problém tzv. sdánlivě nezávislých /seemingly unrelated/ regresí, neboť obě rovnice na sobě závisí prostřednictvím společného parametru σ^2 . Podrobnosti viz např. ZELLNER (1963) a další literatura z této oblasti.

Druhým problémem jsou špatné asymptotické vlastnosti odhadů, které jsou způsobeny tím, že věrohodnostní funkce není pro τ "hladká". Některé výsledky z této oblasti jsou obsaženy v pracích FEDER (1975 a, b).

Úvahy pro modely segmentové regrese jsou analogické, až na to, že se jedná o podmíněný extrém. Některá řešení nalezneme čtenář v práci HUDSON (1966).

Vlastní minimalizace funkce (19) či (20) není numericky příliš obtížná. Algoritmicky nejjednodušší je tzv. kombinatorické vyhledávání /grid search/, ve kterém se postupně zkoumají všechny možné kombinace dělicích bodů τ_j /velmi jednoduše to jde prostřednictvím cyklů/. Pro $p=2$ musíme bez dalších úprav algoritmu provést sice $\binom{T-r-1}{r-1}$ regresí, ale to není pro malá T příliš mnoho. Výpočet lze však podstatně urychlit dvěma, poměrně jednoduchými zásahy:

1. Předem si spočítat tabulku součtu čtverců $Q(t_{j-1}, t_j)$ pro všechny možné oblasti $M_j = \langle t_{j-1}, t_j \rangle$. Tím určíme Q_j pouhým vyhledáním v této tabulce. V případě potřeby lze urychlit výpočty v tabulce použitím rekurentních formulí pro lineární regresní model, které jsou uvedeny např. v práci BROWN, DURBIN, EVANS (1975)
2. Přerušování cyklů, když již částečný součet čtverců $Q = \sum_{j=1}^r Q_j$, $p < r$ překročí již dosažené minimum.

Tyto jednoduché prostředky umožňují řešit úlohu 1 na poměrně malých počítačích velmi rychle a algoritmy jsou velmi přehledné. Podle našeho názoru není nutné používat např. navrhovaných metod dynamického programování.

Další cestou, jak racionalizovat objem výpočtů, je použít věcné omezení na oblasti M_j , které lze použít tam, kde již máme nalezeno $(r-1)$ oblastí.

Hierarchické řešení připouští pouze dělení oblastí. Znamená to, že v prvním kroku budou obě řešení totožná /dělíme celý soubor/, ale ve druhém kroku dělíme v nehierarchickém řešení opět celý soubor na 3 oblasti, zatímco v řešení hierarchickém zkoumáme, zda je lepší rozdělit první či druhou oblast.

Je zřejmé, že je zde možno vytvořit celou řadu dalších modifikací /dělit např. pouze oblast s největším lokálním rozptylem/. V každém případě se jedná o numericky podstatně méně rozsáhlou úlohu.

3. Testování hypotéz.

Nejdůležitější rozhodnutí v částečné a segmentové regresi je, zda platí či neplatí "rozdělený" model. Pro testování hypotéz tohoto typu se používá dvou skupin metod. Podrobněji se, zejména druhou skupinou metod, zabývá ve své monografii HACKL (1980).

3.1. Metody založené na nadmodelu.

Klasickým příkladem testů z této oblasti je nejjednodušší varianta částečné regrese zkoumaná Quandtem.

Alternativy lze popsat v heteroskedastickém modelu jako složenou hypotézu

$$(21) \quad \begin{cases} H_0 : (\beta_1, \sigma_1^2) = (\beta_2, \sigma_2^2) \\ H_1 : (\beta_1, \sigma_1^2) \neq (\beta_2, \sigma_2^2) \end{cases}$$

Zdánlivě nejjednodušší je použít klasického testu věrohodnostním poměrem, kde asymptoticky platí, $-2 \ln \lambda$ má chí-kvadrát rozdělení.

Vzhledem k tomu, že jsou určité problémy s asymptotikou v částečné regresi, nelze však tohoto lákavého přístupu použít /viz např. FEDER (1975b) /. V práci HAWKINS (1977) je odvozen exaktní test tohoto typu pro nejjednodušší případ $p=1$ a homoskedastický model.

Velmi rozšířeným testem je známý Chowův test shody skupiny regresních koeficientů, který můžeme pro homoskedastický případ /tedy pro $\sigma_1^2 = \sigma_2^2 = \sigma^2$ v hypotéze (21) / zapsat ve formě

$$(22) \quad \frac{\frac{(T-p)\hat{\sigma}^2 - T s^2}{T s^2}}{T-2p} \sim F[p, T-2p] ,$$

kde s^2 je průměrovaný odhad σ^2 (17) pro model částečné regrese

a $\hat{\sigma}^2$ je odhad rozptylu náhodných poruch pro nerozdělený model.

/Testy zde definujeme pomocí rozdělení testového kritéria./

Varianta tohoto testu je použitelná i pro testování hypotézy, že počet oblastí je $(r-1)$ oproti alternativě, že jich je r :

$$(23) \quad \frac{\frac{Q^{(r-1)} - Q^{(r)}}{p}}{\frac{Q^{(r)}}{T-pr}} \sim F[p, t-pr] ,$$

kde např. $Q^{(r)}$ je celkový residuální součet čtverců (20) při r dělících bodech.

Určité praktické problémy tohoto lánvého přístupu /navrhovaného např. v práci McGEE, CARLETON (1970) / si ukážeme v aplikační části práce.

Jiný přístup navrhli FARLEY, HINICH (1970). Testování je založeno na nadmodelu

$$(24) \quad y_t = x_t \beta + z_t \delta + \varepsilon_t \quad , t=1, 2, \dots, T \quad ,$$

ve kterém předpokládáme, že známe dělicí bod τ a x_t je vektor vysvětlujících proměnných v časovém okamžiku t .

Vektor umělých proměnných z_t je definován buď jako

$$(25) \quad z_t = \begin{cases} 0 & \text{pro } t < \tau \\ x_t & \text{pro } t \geq \tau \end{cases}$$

pro model částečné regrese, nebo pro model segmentové regrese jako

$$(26) \quad z_t = \begin{cases} 0 & \text{pro } t < \tau \\ x_t - x_\tau & \text{pro } t \geq \tau . \end{cases}$$

Hypotézu o přítomnosti "skoků" v bodě τ , kterou můžeme formulovat ve formě

$$(27) \quad \begin{cases} H_0 : \delta = 0 \quad , \\ H_1 : \delta \neq 0 \quad , \end{cases}$$

můžeme testovat pomocí klasického F-testu o skupině regresních parametrů.

Farley a Hinich v řadě prací provedli i simulační experimenty, ze kterých v podstatě vyplývá, že v případě, že je dělicí bod zhruba v polovině, je vydatnější test Chowův a jinak je lepší jejich test.

Zajímavou praktickou úlohou /zejména v oblasti ekonometrie/ je určit, které z regresních koeficientů se v oblastech M_j významně liší, které jsou "zodpovědné" za rozdělení modelu. Touto problematikou se zabývá /i prakticky/ např. KATZENBERGER (1980).

1.2. Metody založené na analýze residuí.

Touto problematikou se podrobně zabývá zejména HACKL (1980). Testy jsou převážně grafické a jsou analogické testům používaným ve statistické kontrole jakosti.

Prakticky se používají 4 druhy testů:

CUSUM testy, založené na statistikách

$$(28) \quad O_t^{(1)} = \frac{1}{s} \sum_{i=1}^t e_i \quad , t=1,2,\dots,T$$

kde e_i jsou hodnoty residuí /používají se různé druhy /

s^2 je nekreslený odhad rozptylu náhodných poruch

$$(29) \quad s^2 = \frac{1}{T-p} \sum_{i=1}^T e_i^2$$

CUSUM_{SQ} testy jsou založeny na součtech čtverců

$$(30) \quad O_t^{(2)} = \frac{1}{s^2} \sum_{i=1}^t e_i^2 \quad , t=1,2,\dots,T$$

MOSUM testy založené na statistikách

$$(31) \quad M_t^{(1)} = \frac{1}{s} \sum_{i=t-G+1}^t e_i \quad , t=G,\dots,T$$

kde G je počet členů klouzavé části

a konečně

MOSUM_{SQ} testy založené na statistikách

$$(32) \quad M_t^{(2)} = \frac{1}{s^2} \sum_{i=t-G+1}^t e_i^2 \quad , t=G,\dots,T$$

V těchto testech lze sice s určitými těžkostmi použít i klasická residua e_1 získaná metodou nejmenších čtverců pro nerozdělený model /viz např. McGAHE, HARRISON (1979) /, rozumnější je však použít tzv. rekursivních residuí, která mají řadu výhodných vlastností.

Rekursivní residuum w_t je normovaná chyba extrapolace o 1 pozorování dopředu na základě modelu založeného na posledních $(t-1)$ pozorováních .

Je tedy

$$(33) \quad w_t = \frac{y_t - x_t' \cdot \hat{b}_{t-1}}{\sqrt{1 + x_t' (X_{t-1}' X_{t-1})^{-1} x_t}} \quad , \quad t = p, \dots, T \quad .$$

kde X_{t-1} je matice vysvětlujících proměnných v klasickém lineárním regresním modelu pro prvních $(t-1)$ pozorování

$\hat{b}_{t-1} = (X_{t-1}' X_{t-1})^{-1} X_{t-1}' y_{t-1}$ je příslušný odhad

vektoru parametrů metodou nejmenších čtverců.

Tato rekursivní residua jsou nezávislá, mají nulovou střední hodnotu a konstantní rozptyl /samozřejmě za platnosti nulové hypotézy / a v případě klasického lineárního regresního modelu i normální rozdělení.

CUSUM test, založený na rekursivních reziduích navrhl BROWN, DURBIN, EVANS (1975) .

Pro CUSUM statistiku

$$(34) \quad V_t = -\frac{1}{s} \sum_{i=p+1}^t w_i \quad , \quad t = p+1, \dots, T$$

platí v případě nerozděleného modelu

$$(35) \begin{cases} E(V_t) = 0, \\ D(V_t) = t-p, \\ E(V_j, V_k) = \min(j, k) - p, \\ V_t \text{ je asymptoticky normální.} \end{cases}$$

Testování hypotézy (21) /že model není rozdělen proti hypotéze rozdělení modelu/ pak můžeme založit na testování neplatnosti požadavků (35).

Test hypotézy

$$(36) \begin{cases} H_0 : E(V_t) = 0, t=p+1, \dots, T \\ H_1 : E(V_t) \neq 0 \end{cases}$$

můžeme pak skonstruovat na základě toho, že V_t je vlastně náhodná procházka.

H_0 zamítáme v případě, že

$$(37) \quad |\hat{V}_t| > h_\alpha \sqrt{T-p} + 2 h_\alpha \frac{t-p}{\sqrt{T-p}}, \quad t = p+1, \dots, T$$

alespoň pro jedno t .

Zde α je hladina významnosti,

\hat{V}_t jsou napozorované /vypočtené/ hodnoty V_t .

Nejdůležitější hodnoty h_α jsou $h_{0,05} = 0,948$ a $h_{0,01} = 1,143$.

Použijeme-li rekursivní rezidua pro CUSUM_{SQ} test

$$(38) \quad w_t = \frac{1}{2} \sum_{i=p+1}^t w_i^2 = \frac{\sum_{i=p+1}^t w_i^2}{\sum_{i=p+1}^T w_i^2}$$

můžeme využít toho, že

$$(39) \quad w_t \sim \text{Beta} \left[\frac{t-p}{2}, \frac{T-2p}{2} \right]$$

Hypotézu H_0 v testu 36 zamítáme, když

$$(40) \quad |\hat{W}_t| > c + \frac{t-p}{T-p}, \text{ alespoň pro jedno } t,$$

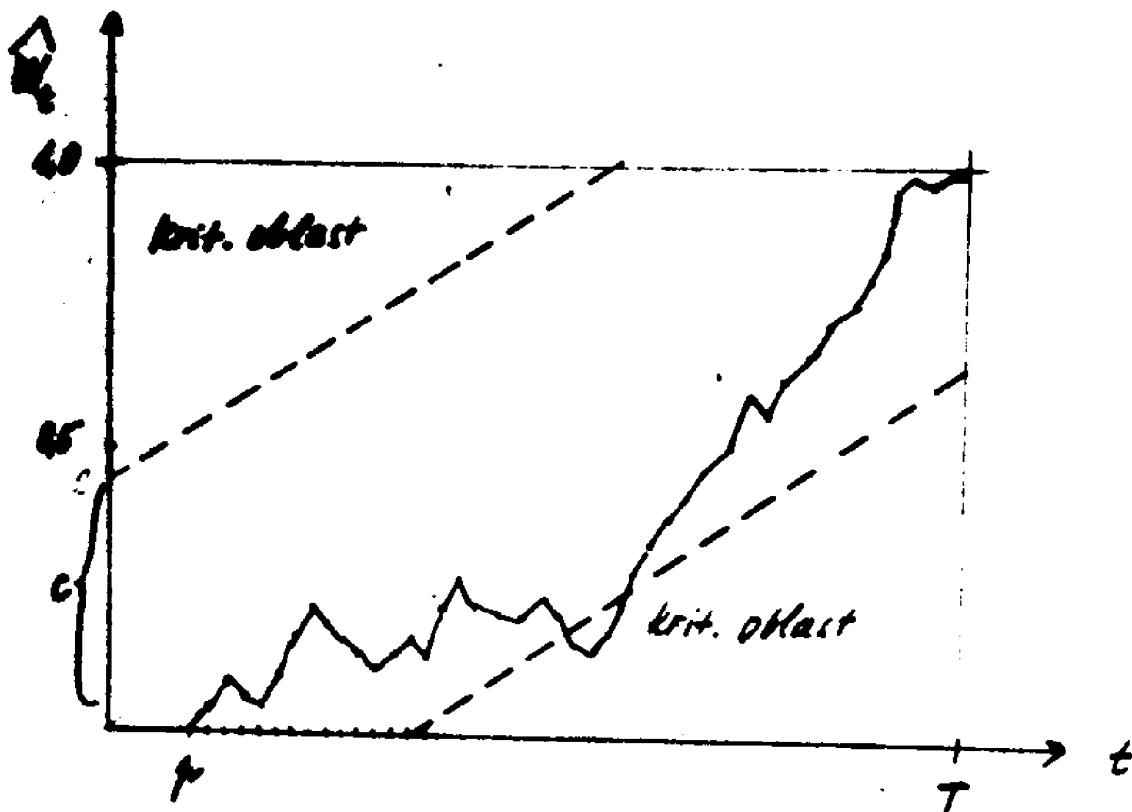
kde hodnotu c nalezneme ve známých tabulkách DURBIN (1969)

se vstupem $n = \frac{T-p}{2} - 1$ a $\frac{\alpha}{2}$, α je hladina významnosti.

Vzhledem k dostupnosti Durbinových tabulek uvádíme zkrácené tabulky pro hladinu významnosti 0,05 /tedy již se správným vstupem/.

n	c	n	c	n	c	n	c	n	c
1	0.47500	11	0.32894	21	0.25622	31	0.21752	41	0.19254
2	.50855	12	.31869	22	.25136	32	.21457	42	0.19050
3	.46702	13	.30935	23	.24679	33	.21173	43	.18852
4	.44641	14	.30081	24	.24254	34	.20901	44	.18661
5	.42174	15	.29296	25	.23835	35	.20639	45	.18475
6	.40045	16	.28570	26	.23445	36	.20387	46	.18295
7	.38294	17	.27897	27	.23074	37	.20144	47	.18120
8	.36697	18	.27270	28	.22721	38	.19910	48	.17950
9	.35277	19	.26685	29	.22383	39	.19684	49	.17785
10	.34022	20	.26137	30	.22061	40	.19465	50	.17624

Tyto testy se zpravidla znázorňují graficky. Do grafů se zaznamenávají průběhy veličin \hat{V}_t či \hat{W}_t a příslušné kritické oblasti. Příklad toho, jak takovýto test v grafickém znázornění vypadá nám ukazuje obrázek číslo 5.



Graf č. 5 - průběh $CUSUM_{SQ}$ testu.

Makolik nám může obrázek pomoci identifikovat i okamžik, kdy ke skoku dochází /jak se někdy z těchto obrásků zdá/ není jasné.

MOSUM testy se zabýval speciálně Hackl. Problém jejich použití je v tom, že obecně neexistuje vůbec žádný návod jak volit délku kloubové části G , proto se jimi zde nebudeme zabývat. Potenciálně jsou tyto testy vhodné tam, kde je "zlomů" více a mohly by se navzájem kompenzovat.

Hackl provedl i řadu simulačních experimentů, ve kterých empiricky skoumal sílu uvedených čtyřech testů. V podstatě z nich vyplývá, že nejlepší je používat $CUSUM_{SQ}$ test.

4. Příklad praktické aplikace.

Pro ilustraci postupu na konkrétním materiálu jsme zvolili dvě makroekonomické časové řady ČSSR a to

A. Společenský produkt v mld. Kčs tzv. srovnatelných cen /tj.

pro léta 1948-60 v cenách k 1.4.1955,

1961-66 -"- 24.4.1960,

1967-75 -"- 1.1.1967,

1976- -"- 1.1.1977 /.

B. Dynamika společenského produktu, což jsou bazické indexy k roku 1948, v nichž jsou cenové skoky eliminovány pomocí tzv. cenových můstků.

Obě časové řady jsou v nezměněné podobě převzaty z publikace VÚSEI Dlouhodobé časové řady, která obsahuje údaje do roku 1980.

Předpokládali jsme, že v jednotlivých oblastech můžeme tyto časové řady modelovat jako časovou řadu s exponenciálním trendem. Pro tento model je adekvátní náhodné poruchy modelovat jako multiplikatívní se střední hodnotou 1 /tedy ve formě indexu/ a konstantním rozptylem. Jako model rozdělení se v tomto případě používá logaritmicko-normální rozdělení s odpovídajícími parametry. Tento model je v oblasti makroekonomických časových řad velmi častý a prakticky vyzkoušený.

Snadno se přesvědčíme, že tato formulace je identická s modelem časové řady s lineárním trendem a normálními náhodnými poruchami pro logaritmy původních hodnot časové řady.

Použitý model tedy je

$$(41) \quad \ln y_t = \alpha_j + \beta_j t + \varepsilon_t, \quad t=1,2,\dots,T$$
$$j=1,2,\dots,r.$$

Pro první časovou řadu se zřejmě jedná o model částečné regrese, zatímco pro druhou časovou řadu se jedná spíše o model segmentové regrese. Obě časové řady jsme však zpracovávali stejně metodami částečné regrese.

Výsledky testování počtu dělicích bodů jsou uvedeny v následujících tabulkách. V tabulce jsou uvedeny zejména výsledky testu (23). Dělicí body lze převést na roky vzorcem $R = t + 1947$.

SPOLEČENSKÝ PRODUKT V MLD SC

počet oblasti	reziduální součet ctvercu $Q(r)$	F-test		delící body						
		(r) proti (r-1)		<i>0 - významnost modelu</i>						
1	0.0980960	2205.099**								
2	0.0725801	5.097*	19							
3	0.0181049	40.619**	13 19							
4	0.0108572	8.344**	2 14 19							
5	0.0044412	16.613**	2 14 19 28							
6	0.0020084	12.718**	2 13 17 19 28							
7	0.0012188	6.154**	2 6 13 17 19 28							
8	0.0007995	4.458*	2 5 7 13 17 19 28							
9	0.0005210	4.007*	2 5 7 12 15 17 19 28							
10	0.0002876	5.276*	2 5 7 12 15 17 19 28 30							

DYNAMIKA SPOL.PRODUKTU V SC

počet oblasti	reziduální součet ctvercu $Q(r)$	F-test		delící body						
		(r) proti (r-1)		<i>významnost modelu</i>						
1	0.1183042	1370.516**	0							
2	0.0223043	62.409**	14							
3	0.0129196	9.806**	15 27							
4	0.0037959	30.044**	2 15 27							
5	0.0020588	9.702**	2 14 16 27							
6	0.0012600	6.656**	2 5 14 16 27							
7	0.0008203	5.091*	2 5 14 16 25 30							
8	0.0004693	6.355**	2 4 7 14 16 25 30							
9	0.0002932	4.505*	2 4 7 12 14 16 25 30							
10	0.0001759	4.333*	2 4 7 12 14 16 22 28 31							

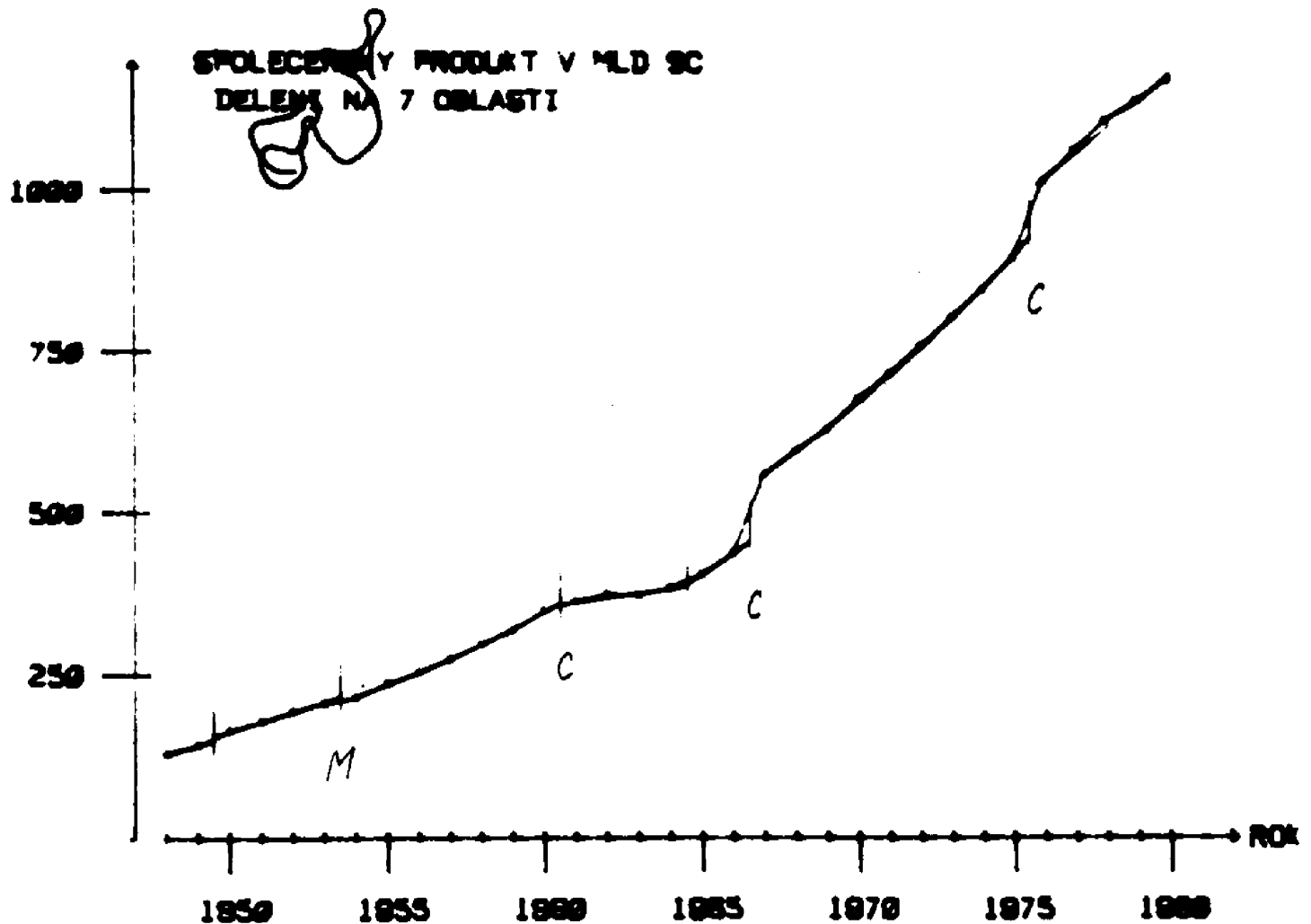
Tabulky poměrně dobře ilustrují problémy vyvstávající s použitím Chowova testu. Na 5% hladině významnosti je každé "rozbití" významné. Dobré výsledky jsme dostali zvýšením hladiny významnosti na 1%, ale i zde se projevuje určitá nestabilita.

Tyto výsledky je tedy nutno konfrontovat s grafickým znázorněním případně s věcnými znalostmi /o tom, co "možné" /.

Na druhé straně je překvapením poměrná stabilita řešení, která jsou do značné míry "hierarchická". To svědčí o tom, že rozbití opravdu existuje a není náhodného původu.

Na základě výsledků uvedených v předchozích tabulkách, a i na základě věcných úvah a grafického porovnání /např. dělení na 7 oblastí s dělením na 8 oblastí/, jsme se rozhodli v obou případech pro model částečné regrese /homoskedastický/ s dělením na 7 oblastí.

Na obrázku číslo 7 jsou znázorněny výsledky pro časovou řadu Společenský produkt ve srovnatelných cenách.

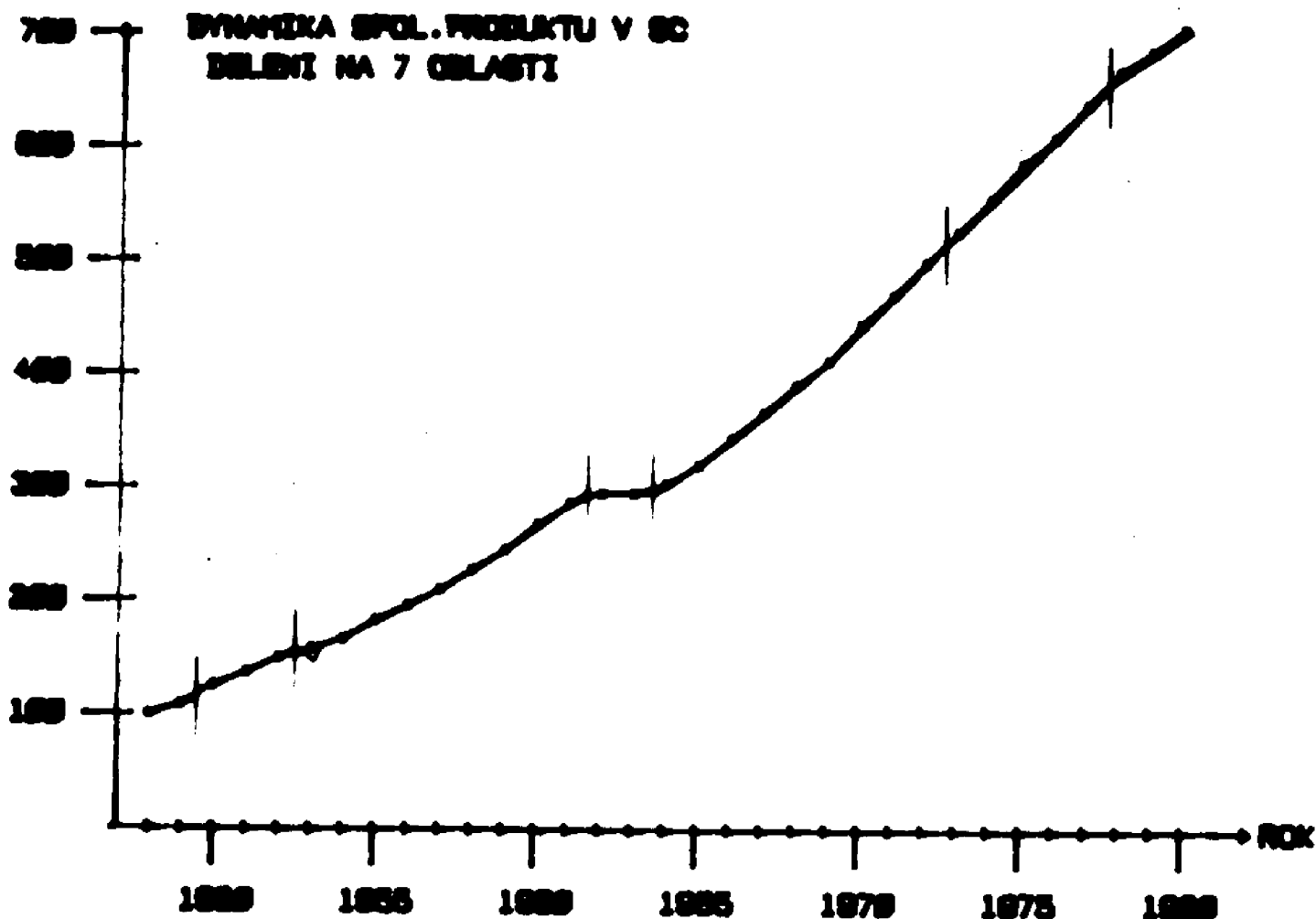


Obrázek č.7

Z obrázku je patrné, že se nám podařilo identifikovat všechny 3 body odpovídající změnám stálých cen /v obrázku označeny C /, dále bod, ve kterém došlo k měnové reformě, která má podobný charakter /označeno M /.

Zbývající dva body jsou rok 1949 a 1964, které lze věcně interpretovat jako body, ve kterých došlo k významné přestavbě národního hospodářství /léta 1949-50/ a rok, ve kterém došlo opět k získání dynamiky po relativní stagnaci let 1961-64.

Výsledky pro časovou řadu Dynamika společenského produktu jsou zobrazeny na obrázku číslo 8.



Obrázek č.8

Ačkoliv se zde jedná o model segmentové regrese, dalo použití metod částečné regrese rozumné výsledky. Návaznost regresních čar existuje, křivka je prakticky spojitá /kromě menších odchylek u prvních dvou dělicích bodů. Je vidět, že pokud se jedná skutečně o model segmentové regrese, nemusíme zavádět předpoklad spojitosti uměle.

! daném případě se již nevyskytují skoky, způsobené změnami cen. Jako zlomy v dynamice se nyní ukazují roky 1949, 1952, 1961, 1963 a 1977 /přesněji řečeno jsou uvedené roky poslední z předchozího období stability - viz obrázek/. Interpretace je zde samozřejmě složitější, ale na druhé straně je posnávací hodnota výsledků vyšší.

Do dalších detailů a svou povahou spíše ekonomických rosborů nechceme na tomto místě zacházet. Zmíníme se pouze o ztrátě dynamiky po roce 1977 /zřejmě v důsledku růstu cen surovin pro čs. hospodářství/ a dvě delší období relativně stabilního vývoje v letech 1953-61 a 1964-72.

V každém případě je z výsledků patrné, že ekonomické časové řady nejsou homogenní, takže pro ně není celá řada statistických metod, které tuto skutečnost nerespektují, použitelná.

5. Závěr.

V této práci jsme chtěli dát čtenáři určitý přehled o metodách částečné a segmentové regrese. Tento text nelze chápat jinak, než pouhý úvod do velmi složité a mnohotvárné problematiky.

Přesto však doufáme, že již na této elementární úrovni lze s těmito metodami prakticky pracovat, a že se nám podařilo ukázat, co lze od nich očekávat.

Zejména doufáme, že poslouží jako inspirace těm, kteří se chtějí něčím zabývat teoreticky, neboť toto téma není podle našeho názoru zpracováno na té úrovni, jako jiné statistické disciplíny.

Literature:

- 1 Brown, R.L.-Durbin, J.-Evans, J.H.: Techniques for testing the constancy of regression relationships over time. JRSS, B37 (1975), s. 149-192.
- 2 Chow, G.C.: Tests of equality between sets of regression coefficients in two linear regressions. Econometrica, 28 (1960), s. 591-605.
- 3 Durbin, J.: Tests for serial correlation in regression analysis based on periodogram of least squares residuals. Biometrika, 56 (1969), s. 1-15.
- 4 Farley, J.U.-Hinich, M.J.: A test for a shifting slope coefficient in a linear model. JASA, 65 (1970), s. 1320-1329.
- 5 Feder, P.L.: On asymptotic distribution theory in segmented regression problems-identified case. Ann. Statist. 3 (1975), s. 49-97.
- 6 Feder, P.L.: The log likelihood ratio on segmented regression. Ann. Statist. 3 (1975), s. 84-97.
- 7 Fisher, F.M.: Tests of equality between sets of coefficients in two linear regressions. An expository note. Econometrica, 38 (1970), s. 361-366.
- 8 Hackl, P.: Testing the constancy of regression models over time. Vandenhoeck-Ruprecht, Göttingen, 1980.
- 9 Hawkins, D.M.: Point estimation of the parameters of piecewise regression models. Appl. Statist. 25 (1976), s. 51-57.
- 10 Hawkins, D.M.: A note on continuous and discontinuous segmented regression. Technometrics, 22 (1980), s. 443-444.
- 11 Hinkley, D.V.: Inference in two-phase regression. JASA, 55 (1971), s. 324-330.
- 12 Katzenberger, W.: Test auf Gleichheit von Regressionskoeffizienten: Einige Erweiterungen. Statistische Hefte, 22 (1981), s. 25-39.
- 13 Leruan, P.M.: Fitting segmented regression models by grid search. Appl. Statist. 29 (1980), s. 77-84.
- 14 McCabe, B.P.M.-Harrison, M.J.: Test of the constancy of regression relationships over time using least squares residuals. Appl. Statist. 29 (1980), s. 142-148.
- 15 McGee, V.E.-Carleton, W.T.: Piecewise regression. JASA 65 (1970), s. 1109-1124.

- 16 Oberhofer, W.: Die Nichtkonsistenz der M.L. Schätzer in "switching regression" Problem. *Metrika*, 27 (1980), s. 1-13.
- 17 Poirier, D.J.: The econometrics of structural change. North Holland, Amsterdam, 1976.
Existuje též ruský překlad Puarje, D.: *Ekonometrija strukturnych izmerenij*. *Finansy i statistika*, Moskva, 1981.
- 18 Quandt, R.E.: The estimation of the parameters of a linear regression system obeying two separate regimes. *JASA*, 53 (1958), s. 873-880.
- 19 Quandt, R.E.: Tests of hypothesis that a linear regression system obeys two separate regimes. *JASA*, 55 (1960), s. 324-330.
- 20 Quandt, R.E.: A new approach to estimating switching regressions. *JASA*, 67 (1972), s. 306-310.
- 21 Toyoda, T.: Use of the Chow test under heteroscedasticity. *Econometrica*, 42 (1974), s. 601-608.
- 22 Zellner, A.: An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *JASA*, 57 (1962), s. 348-362.

- 11a Hudson, D.J.: Fitting segmented curves, whose join points have to be estimated. *JASA*, 61 (1966), s. 1097-1129.